

Dynamic Large Spatial Covariance Matrix Estimation in Application to Semiparametric Model Construction via Variable Clustering: the SCE approach

Song Song*

June 24, 2011

Abstract

To better understand the spatial structure of large panels of economic and financial time series and provide a guideline for constructing semiparametric models, this paper first considers estimating a large spatial covariance matrix of the generalized m -dependent and β -mixing time series (with J variables and T observations) by hard thresholding regularization as long as $\log J \mathcal{X}^*(\mathcal{T})/T = o(1)$ (the former scheme with some time dependence measure $\mathcal{X}^*(\mathcal{T})$) or $\log J/T = o(1)$ (the latter scheme with the mixing coefficient $\beta_{mix} = \mathcal{O}\{(J^{2+\delta'} \sqrt{\log JT})^{-1}\}, \delta' > 0$). We quantify the interplay between the estimators' consistency rate and the time dependence level, discuss an intuitive resampling scheme for threshold selection, and also prove a general cross-validation result justifying this. Given a consistently estimated covariance (correlation) matrix, by utilizing its natural links with graphical models and semiparametrics, after “screening” the (explanatory) variables, we implement a novel forward (and backward) label permutation procedure to cluster the “relevant” variables and construct the corresponding semiparametric model, which is further estimated by the groupwise dimension reduction method with sign constraints. We call this the SCE (screen - cluster - estimate) approach for modeling high dimensional data with complex spatial structure. Finally we apply this method to study the spatial structure of large panels of economic and financial time series and find the proper semiparametric structure for estimating the consumer price index (CPI) to illustrate its superiority over the linear models.

Keywords: Time Series, Covariance Estimation, Regularization, Sparsity, Thresholding, Semiparametrics, Graphical Model, Variable Clustering

JEL classification: C13, C14, C32, E30, G10

*University of California, Berkeley. Email: songsong@stat.berkeley.edu

1 Introduction

1.1 Large Spatial Covariance Matrix

Recent breakthroughs in technology have created an urgent need for high-dimensional data analysis tools. Examples include economic and financial time series, genetic data, brain imaging, spectroscopic imaging, climate data and many others. To model high dimensional data, especially large panels of economic and financial time series as our focus here, it is very important to begin with understanding the “spatial” structure (over the space of variables instead of from a geographic point of view; also used in future for convenience) instead of simply assuming any specific type of parametric (e.g. linear) model first. Estimation of large spatial covariance matrix plays a fundamental role here since it can indicate a predictive relationship that can be exploited in practice. It is also very important in numerous other areas of economics and finance, including but not limited to handling heteroscedasticity of high dimensional econometric models, risk management of large portfolios, setting confidence intervals (or interval forecasts) on linear functions of the means of the components, variable grouping via graphs, dimension reduction by principal component analysis (PCA) and classification by linear or quadratic discriminant analysis (LDA and QDA). In recent years, many application areas where these tools are used have dealt with very high-dimensional datasets with relatively small sample size, e.g. the typically low frequency macroeconomic data.

It is well known by now that the empirical covariance matrix for samples of size T from a J -variate Gaussian distribution, $N_J(\mu, \Sigma_J)$ is not a good estimator of the population covariance if J is large. If $J/T \rightarrow c \in (0, 1)$ and the covariance matrix $\Sigma_J = I$ (the identity), then the empirical distribution of the eigenvalues of the sample covariance matrix $\hat{\Sigma}_J$ follow the Marčenko-Pastur Law (Marčenko and Pastur (1967)) and the eigenvalues are supported on $((1 - \sqrt{c})^2, (1 + \sqrt{c})^2)$. Thus, the larger J/T is, the more spread out the eigenvalues are.

Therefore, alternative estimators for large covariance matrices have attracted a lot of attention recently. Two broad classes of covariance estimators have emerged. One is to remedy the sample covariance matrix and construct a better estimate by using approaches such as banding, tapering and thresholding. The other is to reduce dimensionality by imposing some structure on the data such as factor models, Fan et al. (2008). Among the first class, regularizing the covariance matrix by banding or tapering relies on a natural ordering among variables and assumes that variables far apart in the ordering are only weakly correlated, Wu and Pourahmadi (2003), Bickel and Levina (2008b), Cai and Zhou (2011) among others. However, there are many applications, such as large panels of macroeconomic and financial time series, gene expression arrays and other spatial data, where there is

no total ordering on the plane and no defined notion of distance among variables at all. These existing applications require estimators to be invariant under variable permutations such as regularizing the covariance matrix by thresholding, El Karoui (2008) and Bickel and Levina (2008a). In this paper, we consider thresholding of the sample spatial covariance matrix for high dimensional time series, which extends the existing work from the iid to the dependent scenarios. Under the time series setup, a very important question to ask is: how the time dependence will affect the estimate’s consistency? This is the first question this paper is going to answer.

For time series, there have been two recent works by Bickel and Gel (2011) and Xiao and Wu (2011) about banding and tapering the large autocovariance matrices for univariate time series. But our goal here is to better understand the spatial structure of high dimensional time series, so we need a consistent estimate of the large spatial covariance matrix, especially under a mixture of serial correlation (temporal dynamics), high dimensional (spatial) dependence structure and moderate sample size (relative to dimensionality).

1.2 Relation with Semiparametric Model Construction

As mentioned at the very beginning, when the spatial structure of the high dimensional data (time series) is complex, instead of simply assuming any specific type of parametric (e.g. linear) model first, we could adopt the flexible nonparametric approach. Due to the “curse of dimensionality” disadvantage of full nonparametrics, various semiparametric models have been considered to maintain flexibility in modeling while attempting to deal with the “curse of dimensionality” problem. However, most of the prior semiparametric works were carried out under some (prefixed) specific classes of semiparametric models without discussing which ones might be closer to the actual data structure. More specifically, to model some dependent variable y (or x_J) using explanatory variables x_1, x_2, \dots, x_{J-1} (very large $J - 1$), they might suggest the following high dimensional single index (Huang et al. (2010)) or additive models (Meier et al. (2009), Ravikumar et al. (2009)) first and then perform various variable selection techniques to eliminate some x ’s to avoid overfitting.

- $E(y) = g(x_1\beta_1 + x_2\beta_2 + x_3\beta_3 + x_4\beta_4 + \dots + x_{J-1}\beta_{J-1})$, where g is an unknown univariate link function, and $\beta_1, \beta_2, \beta_3, \dots, \beta_{J-1}$ are unknown parameters that belong to the parameter space.
- $E(y) = g_1(x_1) + g_2(x_2) + g_3(x_3) + \dots + g_{J-1}(x_{J-1})$, where $g_1, g_2, g_3, \dots, g_{J-1}$ are the unknown functions to be estimated nonparametrically.

This approach encounters limitations from the following three perspectives. *First*, when the dimensionality $J - 1 \rightarrow \infty$, the prefixed assumption itself becomes more and more questionable.

Is the single index model or the additive one closer to the actual data structure? Or maybe some other type of semiparametric structures is more suitable? We do not know. And this becomes more challenging when the sample size T is small (with respect to dimensionality).

Second, this - prefixing some specific semiparametric classes first and then selecting variables accordingly - approach is also challenged by another character of high dimensional economic and financial time series: strong spatial dependence (near-collinearity). Under near-collinearity, we expect variable selection to be unstable and very sensitive to minor perturbation of the data. In this sense, we do not expect variable selection to provide results that lead to clearer economic interpretation than principal components or ridge regression. This is actually due to the fact that although compared with the information criteria based L_0 and ridge regression type L_2 regularization methods, the Lasso type L_1 variable selection techniques (Tibshirani (1996)) could deal with large J and require weaker assumptions on the design matrix x (composed of x_1, \dots, x_{J-1}), it still requires the following (as one of many similar requirements) *restricted eigenvalue (RE)* assumptions from Bickel et al. (2009): *there exists a positive number $\kappa = \kappa(s)$ such that*

$$\min \left\{ \frac{|x^\top \Delta|_2}{\sqrt{T}|\Delta_{\mathcal{R}}|_2} : |\mathcal{R}| \leq s, \Delta \in \mathbb{R}^{J-1} \setminus \{0\}, \|\Delta_{\mathcal{R}^c}\|_1 \leq 3 \|\Delta_{\mathcal{R}}\|_1 \right\} \geq \kappa,$$

where $|\mathcal{R}|$ denotes the cardinality of the set \mathcal{R} , \mathcal{R}^c denotes the complement of the set of indices \mathcal{R} , and $\Delta_{\mathcal{R}}$ denotes the vector formed by the coordinates of the vector Δ w.r.t. the index set \mathcal{R} . It is essentially a restriction on the eigenvalues of the Gram matrix $\Psi_T = x^\top x / T$ as a function of sparsity s . To see this, recall the definitions of *restricted eigenvalue* and *restricted correlation* in Bickel et al. (2009):

$$\begin{aligned} \psi_{\min}(u) &= \min_{z \in \mathbb{R}^{J-1}: 1 \leq \mathcal{M}(z) \leq u} \frac{z^\top \Psi_T z}{|z|_2^2}, \quad 1 \leq z \leq J-1, \\ \psi_{\max}(u) &= \max_{z \in \mathbb{R}^{J-1}: 1 \leq \mathcal{M}(z) \leq u} \frac{z^\top \Psi_T z}{|z|_2^2}, \quad 1 \leq z \leq J-1, \\ \psi_{m_1, m_2} &= \max \left\{ \frac{f_1^\top x_{I_1}^\top x_{I_2} f_2}{T|f_1|_2|f_2|_2} : I_1 \cap I_2 = \emptyset, |I_i| \leq m_i, f_i \in \mathbb{R}^{I_i} \setminus \{0\}, i = 1, 2 \right\}, \end{aligned}$$

where $|I_i|$ denotes the cardinality of I_i and x_{I_i} is the $T \times |I_i|$ submatrix of x obtained by removing from x the columns that do not correspond to the indices in I_i . Lemma 4.1 in Bickel et al. (2009) shows that if the *restricted eigenvalue* of the Gram matrix Ψ_T satisfies $\psi_{\min}(2s) > 3\psi_{s, 2s}$ for some integer $1 \leq s \leq (J-1)/2$, Assumption RE holds. Under this condition, the Lasso type estimate's various oracle inequalities could be derived, e.g. Bickel et al. (2009), where the upper bounds typically negatively depend on κ . From an economic point of view, this in fact requires that the dependence can not be too strong, which, unfortunately, is often unsatisfied for large panels of macroeconomic and financial data.

Third, when the proposed high dimensional semiparametric model has a complex structure, finding a proper penalty term and the corresponding estimation method for variable selection in general might be very difficult, since, ideally, the penalty should depend not only on the coefficients, but also on the (shapes of the) *unknown* nonparametric link functions. Several examples of regularizing high dimensional semiparametric models could be found in Chapter 5 and 8 of Bühlmann and van de Geer (2011), Ravikumar et al. (2009) among others.

To this end, developing a *specific model free* high dimensional spatial structure *first* and then constructing the right class of semiparametric models seems important. Specifically speaking, given x_1, x_2, \dots, x_{J-1} and y (or x_J), we try to find the index sets $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_S$ (possibly with overlapping elements) such that y could be well approximated by:

$$\sum_{s=1}^S g_s \left(\sum_{l=1}^{|\mathcal{A}_s|} \beta_{sl} x_{l \in \mathcal{A}_s} \right) \stackrel{\text{def}}{=} \sum_{s=1}^S g_s \left(\beta_s^\top x_{\mathcal{A}_s} \right), \quad (1)$$

where

- $|\cdot|$ denotes the cardinality of the set \cdot ; S is the number of index sets $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_S$ and also the number of the *unknown* univariate nonparametric link functions g_1, \dots, g_S ;
- $x_{\mathcal{A}_s} \stackrel{\text{def}}{=} (x_l, l \in \mathcal{A}_s)$ is a vector of regressors w.r.t. the index set \mathcal{A}_s ; $\beta_{sl}, 1 \leq s \leq S, 1 \leq l \leq |\mathcal{A}_s|$ are the *unknown* parameters in the parametric space; $\beta_s = (\beta_{s1}, \dots, \beta_{s|\mathcal{A}_s|})$;
- $\forall j \neq l, s \neq t, x_j \in \mathcal{A}_s, x_l \in \mathcal{A}_t, x_j$ and x_l are (conditionally) independent given other x 's.

If $K \stackrel{\text{def}}{=} |\mathcal{A}_1 \cup \dots \cup \mathcal{A}_S| \ll J - 1$ (although K is still possibly not moderate), we could strike a balance between dimension reduction and flexibility of modeling. Model (1) is very general and includes the single index model if $S = 1$ (Ichimura (1993)), the additive model if $|\mathcal{A}_1| = |\mathcal{A}_2| = \dots = |\mathcal{A}_S| = 1$ (Hastie and Tibshirani (1990)), the partial linear model if $S = 2, g_1$ is the identity function and $|\mathcal{A}_2| = 1$ (Speckman (1988)) and the partial linear single index model if $S = 2$ and g_1 is the identity function (Ahn and Powell (1993), Carroll et al. (1995), Yu and Ruppert (2002)). Model (1) could also be viewed as an extension of the multiple index model (Stoker (1986), Ichimura and Lee (1991), Horowitz (1998), Xia (2008)) and can be further generalized if the RHS is $\mu\{E(y)\}$ where μ is some known link function. As also considered by Li et al. (2010), if $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_S$ are disjoint (no overlapping elements), then for each group of variables $x_{\mathcal{A}_s}$, we could say g_s denotes (the only) one index. Thus according to Li et al. (2010), model (1) is identifiable as every subspace of every group $x_{\mathcal{A}_s}$ is identifiable and could be solved efficiently by the grouping dimension reduction method in Li et al. (2010), where they primarily assume that the grouping information is available. An immediate

question is that given x_1, \dots, x_{J-1} ($J - 1 \rightarrow \infty$), how can we extract S groups of “relevant” x ’s with corresponding index sets $\mathcal{A}_1, \dots, \mathcal{A}_S$ and $|\mathcal{A}_1 \cup \dots \cup \mathcal{A}_S| \ll J - 1$? This is the second question this paper is going to answer. From now on, we mainly study the case where $\mathcal{A}_1, \mathcal{A}_2, \dots, \mathcal{A}_S$ are disjoint, although in Section 4 we will also present the method generating overlapping index sets.

Before moving on, let us study the differences among various semiparametric models from the graphical point of view. If we use a vertex in the graph to represent a relevant variable, a solid edge in a “block” to represent linear relationship among variables inside, a bandy edge (connecting a “block” with the dependent variable y) to represent a nonparametric link function, a crossed vertex to represent an “unrelated” ones, then we can visualize different semiparametric models through corresponding graphs. For instance, we can get Figure 1 (left) for the single index model; Figure 1 (right) for the additive model; Figure 2 (left) for the (more general) multiple index model, among many others. As we can see, the underlying difference among various semiparametric models is where to allocate the nonparametric link function and linearity through clustering variables. Consequently, assuming that all the variables have been included (complete graph), if we can find the corresponding type of graphs, we can construct the right class of semiparametric models. Sparse concentration matrices are of special interest in graphical models because zero partial correlations help establish independence and conditional independence relations in the context of graphical models and thus imply a graphical structure. For example, if we have a sparse covariance matrix for y, x_1, \dots, x_9 as the one in Figure 2 (right), we know that x_1, \dots, x_6 are “relevant” to y , and due to the “block” structure w.r.t. x_1, x_2, x_3 and x_4, x_5 , we can construct the following class of semiparametric models as a specific case of (1):

$$E(y) = g_1(x_1\beta_1 + x_2\beta_2 + x_3\beta_3) + g_2(x_4\beta_4 + x_5\beta_5) + g_3(x_6\beta_6). \quad (2)$$

Now we have found the links among semiparametrics, graphical models and sparse large spatial covariance matrix. Thus consistently estimating the large sparse covariance matrix first and clustering the (explanatory) variables (or forming a block diagonal structure for the corresponding partition of the covariance matrix) are the key focuses. In this article, we assume that the grouping structure (or the corresponding covariance matrix) and parametric coefficients β s are both time invariant to simply the study.

Another related and potential application of clustering variables comes from group regularization (e.g. group Lasso, Yuan and Lin (2006)) in the modern sparsity analysis. Huang and Zhang (2009) show that, if the underlying structure is strongly group-sparse, group Lasso is more robust to noise due to the stability associated with group structure and thus requires a smaller sample size to meet the sparse eigenvalue condition required in modern sparsity analysis. However, other than the situations, e.g. multi-task learning, where we have clear background knowledge about how to group variables, in

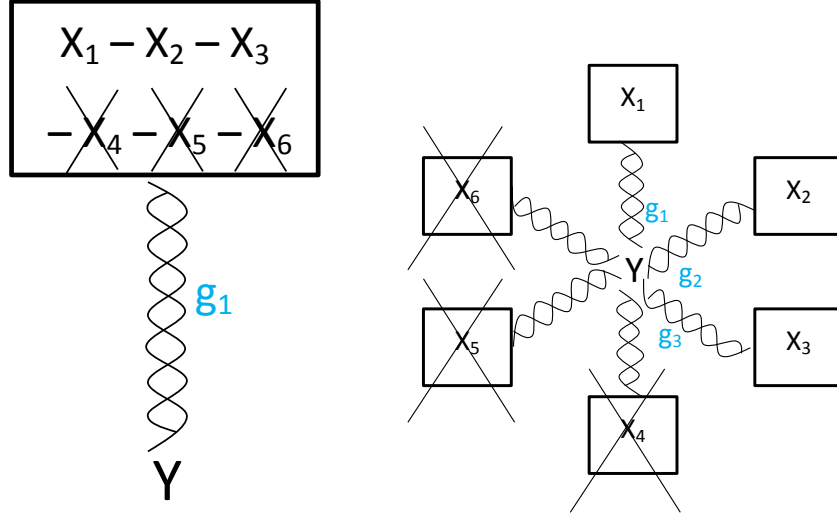


Figure 1: Left: $E(y) = g(x_1\beta_1 + x_2\beta_2 + x_3\beta_3)$, where g is an unknown univariate link function, and $\beta_1, \beta_2, \beta_3$ are unknown indices which belong to the parameter space. Right: $E(y) = g_1(x_1) + g_2(x_2) + g_3(x_3)$, where g_1, g_2 and g_3 are the unknown functions to be estimated nonparametrically.

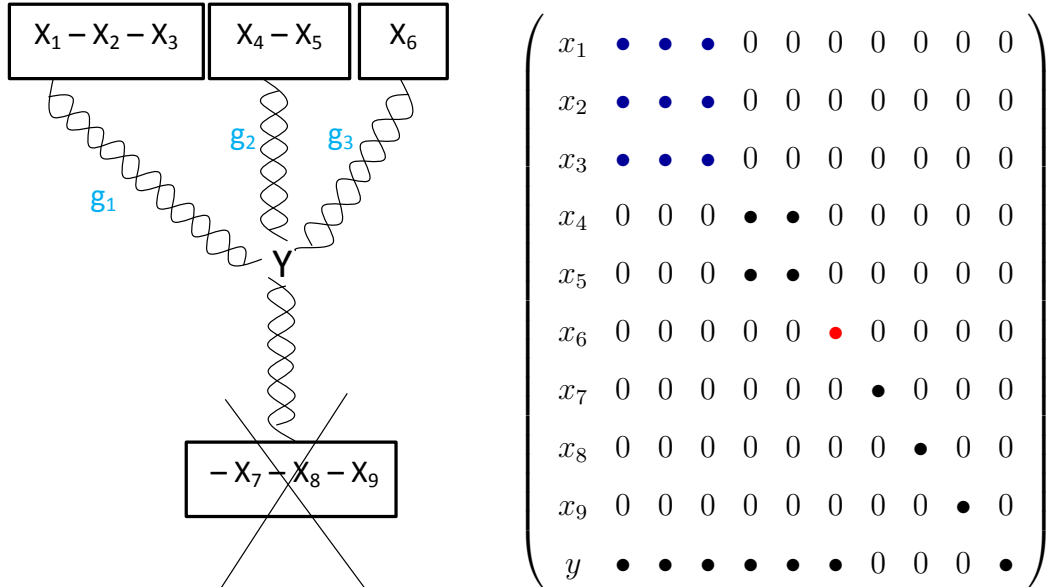


Figure 2: Left: $E(y) = g_1(x_1\beta_1 + x_2\beta_2 + x_3\beta_3) + g_2(x_4\beta_4 + x_5\beta_5) + g_3(x_6\beta_6)$, where g_1, g_2 and g_3 are the unknown functions to be estimated nonparametrically. Right: a sample of block diagonal structure after label permutation to the regularized large spatial covariance matrix.

general, it is hard to tell how to properly group the variables to make use of group regularization. An example could be found in a paper in preparation with Bickel, Song and Bickel (2011), where they discuss three types of estimates for large vector auto regression w.r.t. different grouping methods. To this end, proper “grouping” of the variables is also significant.

For the semiparametric modeling in econometrics, people usually “group” the variables in a “rule of thumb” way. For example, to model the consumer price index (CPI - all items), they might subjectively group the variables “CPI - apparel & upkeep; transportation; medical care; commodities; durables; services” in the first group, “CPI - all items less food; all items less shelter; all items less medical care” in the second group; “Producer Price Index (PPI) - Finished Goods; Finished Consumer Goods; Intermed Mat. Supplies & Components; Crude Materials” in the third group, “Implicit Price Deflator (of Personal Consumption Expenditures) PCE - all items; durables; nondurables; services” in the fourth group, all other variables in the last group. Is this way of grouping closest to the actual data structure? Why not put “CPI - Durables; PCE - Durables” in one group and “CPI - Services; PCE - Services” in another group? We are going to provide a procedure of grouping these variables from a data-driving approach.

In summary, the novelty of this article lies in the following two aspects. *First*, under the high dimensional time series situation, we show consistency (and the explicit rate of convergence) of the threshold estimator in the operator norm, uniformly over the class of matrices that satisfy our notion of sparsity as long as $\log J \mathcal{X}^*(\mathcal{T})/T = o(1)$ (for the generalized m -dependent time series; the meaning of $\mathcal{X}^*(\mathcal{T})$ is presented later) or $\log J/T = o(1)$ (for the β -mixing process with the mixing coefficient $\beta_{mix} = \mathcal{O}\{(J^{2+\delta'} \sqrt{\log JT})^{-1}\}$, $\delta' > 0$). Furthermore, we quantify the interplay between the estimators’ consistency rate and the time dependence level, which is novel in this context. There are various arguments showing that convergence in the operator norm implies convergence of eigenvalues of eigenvectors, El Karoui (2008) and Bickel and Levina (2008a), so this norm is particularly appropriate for various applications. We also discuss an intuitive resampling scheme for threshold selection for high dimensional time series, and prove a general cross-validation result that justifies this approach. *Second*, we propose a SCE (screen - cluster - estimate) approach for modeling high dimensional data with complex spatial structure. Specifically, given a consistently estimated large spatial covariance (correlation) matrix, by utilizing its natural links with graphical models and semiparametrics and using the correlation (or covariance for the standardized observations) between variables as a measure of similarity, after “screening” the (explanatory) variables, we propose a novel forward (and backward) label permutation procedure to cluster the “relevant” (explanatory) variables (or to form a block diagonal structure for the regularized large spatial matrix) and construct the corresponding

semiparametric model, which is further estimated by the groupwise dimension reduction method (Li et al. (2010)) with sign constraints.

It is noteworthy that the “screening” in Step 1, “clustering” in Step 2, and the “sign constraints” in Step 3 here are crucial for applying the groupwise dimension reduction method of Li et al. (2010) in the high dimensional situation. *First*, their method requires the use of the high dimensional kernel function, which faces some limitations when $J \gg T$. The Step 1 here help reduce the dimensionality from $J (\gg T)$ to a more manageable level. *Second*, they primarily assume that the grouping information is available from the background knowledge, which is often not available from the typically (spatially) unordered high dimensional data sets. Although they also proposed an information criterion based grouping method, this - “trying” many different combinations of grouping - approach is very computationally intensive and less practical. The Step 2 here provides this grouping information from a data driven approach with feasible computation. *Third*, without adding the “sign constraints”, the signs of the estimated parametric coefficients might violate the economic laws (details presented in Section 5). Overall, together with Li et al. (2010)’s very timely and stimulating work, we provide an *integrated* approach for modeling high dimensional data with complex spatial structure.

The rest of the article is organized as follows. In the next section, we present the main notations of the thresholding estimator. The estimates’ properties are presented in Section 3. In Section 4 we state the details of the SCE procedure and in Section 5 apply it to study the spatial structure of large panels of macroeconomic and financial times series and find the proper semiparametric structure for estimating the consumer price index (CPI). Section 6 contains concluding remarks with a brief discussion. All technical proofs are sketched in the appendix.

2 Dynamic Large Spatial Covariance Matrix Estimation

We start by setting up notations and corresponding concepts for covariance matrix Σ , which are mostly from Bickel and Levina (2008b) and Bickel and Levina (2008a). We write $\lambda_{\max}(\Sigma) = \lambda_1(\Sigma) \geq \dots \geq \lambda_J(\Sigma) = \lambda_{\min}(\Sigma)$ for the eigenvalues of a matrix Σ . Following the notations of Bickel and Levina (2008b) and Bickel and Levina (2008a), we define that, for any $0 \leq r, s \leq \infty$ and a $J \times J$ matrix Σ , $\|\Sigma\|_{(r,s)} \stackrel{\text{def}}{=} \sup\{\|\Sigma x\|_s : \|x\|_r = 1\}$, where $\|x\|_r^r = \sum_{j=1}^J |x_j|^r$. In particular, we write $\|\Sigma\| = \|\Sigma\|_{(2,2)} = \max_{1 \leq j \leq J} |\lambda_j(\Sigma)|$, which is the operator norm for a symmetric matrix. We also use the Frobenius matrix norm, $\|\Sigma\|_F^2 = \sum_{i,j} \sigma_{ij}^2 = \text{tr}(\Sigma \Sigma^\top)$. Dividing it by a factor J brings $\|\Sigma\|_F^2/J$, which is the average of a set of eigenvalues, while the operator norm $\|\Sigma\|_{(2,2)}$ means the maximum of

the same set of eigenvalues. Bickel and Levina (2008a) defines the thresholding operator by

$$T_s(\Sigma) \stackrel{\text{def}}{=} [m_{ij} \mathbf{1}(|m_{ij}| \geq s)],$$

which we refer to as Σ thresholded at s . Notice that T_s preserves symmetry; it is invariant under permutations of variable labels; and if $\|T_s - T_0\| \leq \varepsilon$ and $\lambda_{\min}(\Sigma) > \varepsilon$, it preserves positive definiteness.

We study the properties of the following uniformity class of covariance matrices invariant under permutations

$$\mathcal{U}_\tau(q, c_0(J), M) \stackrel{\text{def}}{=} \{\Sigma : \sigma_{ii} \leq M, \sum_{j=1}^J |\sigma_{ij}|^q \leq c_0(J), \forall i\}, \quad 0 \leq q < 1.$$

We will mainly write c_0 for $c_0(J)$ in the future. Suppose that we observe T J -dimensional observations X_1, \dots, X_T with $\mathbb{E} X = 0$ (without loss of generality), and $\mathbb{E}(XX^\top) = \Sigma$, which is independent of t . We consider the sample covariance matrix by

$$\hat{\Sigma} \stackrel{\text{def}}{=} T^{-1} \sum_{t=1}^T (X_t - \bar{X})(X_t - \bar{X})^\top \stackrel{\text{def}}{=} [\hat{\sigma}_{ij}], \quad (3)$$

with $\bar{X} = T^{-1} \sum_{t=1}^T X_t$.

Let us first recall the fractional cover theory based definition, which was introduced by Janson (2004) and can be viewed as a generalization of m -dependency. Given a set \mathcal{T} and random variables $V_t, t \in \mathcal{T}$, we say:

- A subset \mathcal{T}' of \mathcal{T} is *independent* if the corresponding random variables $\{V_t\}_{t \in \mathcal{T}'}$ are independent.
- A family $\{\mathcal{T}_j\}_j$ of subsets of \mathcal{T} is a *cover* of \mathcal{T} if $\bigcup_j \mathcal{T}_j = \mathcal{T}$.
- A family $\{(\mathcal{T}_j, w_j)\}_j$ of pairs (\mathcal{T}_j, w_j) , where $\mathcal{T}_j \subseteq \mathcal{T}$ and $w_j \in [0, 1]$ is a *fractional cover* of \mathcal{T} if $\sum_j w_j \mathbf{1}_{\mathcal{T}_j} \geq \mathbf{1}_{\mathcal{T}}$, i.e. $\sum_{j: t \in \mathcal{T}_j} w_j \geq 1$ for each $t \in \mathcal{T}$.
- A (fractional) cover is *proper* if each set \mathcal{T}_j in it is independent.
- $\mathcal{X}(\mathcal{T})$ is the size of the smallest proper cover of \mathcal{T} , i.e. the smallest m such that \mathcal{T} is the union of m independent subsets.
- $\mathcal{X}^*(\mathcal{T})$ is the minimum of $\sum_j w_j$ over all proper fractional covers $\{(\mathcal{T}_j, w_j)\}_j$.

Notice that, in spirit of these notations, $\mathcal{X}(\mathcal{T})$ and $\mathcal{X}^*(\mathcal{T})$ depend not only on \mathcal{T} but also on the family $\{V_t\}_{t \in \mathcal{T}}$. Further note that $\mathcal{X}^*(\mathcal{T}) \geq 1$ (unless $\mathcal{T} = \emptyset$) and that $\mathcal{X}^*(\mathcal{T}) = 1$ if and only if the variables $V_t, t \in \mathcal{T}$ are independent, i.e. $\mathcal{X}^*(\mathcal{T})$ is a measure of the dependence structure of $\{V_t\}_{t \in \mathcal{T}}$.

For example, if V_t only depends on V_{t-1}, \dots, V_{t-k} but is independent of all $\{V_s\}_{s < t-k}$, we will have $k+1$ independent sets:

$$\begin{aligned}\mathcal{T}_1 &= \{V_1, V_{(k+1)+1}, V_{2(k+1)+1}, \dots\}, \\ \mathcal{T}_2 &= \{V_2, V_{(k+1)+2}, V_{2(k+1)+2}, \dots\}, \\ &\dots \\ \mathcal{T}_{k+1} &= \{V_{k+1}, V_{(k+1)+(k+1)}, V_{2(k+1)+(k+1)}, \dots\},\end{aligned}$$

s.t. $\bigcup_{j=1}^{k+1} \mathcal{T}_j = \mathcal{T}$. So $\mathcal{X}^*(\mathcal{T}) = k+1$ (if $k+1 < T$).

Besides the generalized m -dependent process, we are also going to consider the β -mixing process, which is related to the underlying measures of dependence between σ -fields. More precisely, let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and \mathcal{U}, \mathcal{V} be two sub σ -algebras of \mathcal{A} , the β -mixing coefficient $\beta(\mathcal{U}, \mathcal{V}) = \mathbb{E} \operatorname{esssup}\{|\mathbb{P}(V|\mathcal{U} - \mathbb{P}(V))|; V \in \mathcal{V}\}$ be a measure of dependence between \mathcal{U} and \mathcal{V} , which has been defined by Kolmogorov and first appeared in the paper by Volkonskii and Rozanov (1959). By its definition, the closer to 0 β is, the more independent the time series is. For examples of the β -mixing process, we refer to Doukhan (1994). Through this article, we use β_{mix} to denote the β -mixing coefficient for notational convenience.

3 Estimates' Properties

We have the following two results which parallel those in Bickel and Levina (2008b) and Bickel and Levina (2008a).

3.1 Interplay Between Consistency Rate and Time Dependence Level

THEOREM 3.1 (Dependence level affects consistency?) *Suppose for all i, j , $|X_{ti}X_{tj}| \stackrel{\text{def}}{=} |V_t| \leq M_t$ holds with a high probability and $\sum_{t=1}^T M_t^2/T$ is bounded by some constant C' . Then, uniformly on $\mathcal{U}_\tau(q, c_0(J), M)$, for sufficiently large M' also depending on C' , if*

$$s_T = M'(C') \sqrt{\frac{\log J \mathcal{X}^*(\mathcal{T})}{T}}$$

and $\log J \mathcal{X}^*(\mathcal{T})/T = o(1)$, then

$$\begin{aligned}\|T_{s_T}(\hat{\Sigma}) - \Sigma\| &= \mathcal{O}_P \left[c_0(J) \left\{ \frac{\log J \mathcal{X}^*(\mathcal{T})}{T} \right\}^{(1-q)/2} \right] \\ J^{-1} \|T_{s_T}(\hat{\Sigma}) - \Sigma\|_F^2 &= \mathcal{O}_P \left[c_0(J) \left\{ \frac{\log J \mathcal{X}^*(\mathcal{T})}{T} \right\}^{1-q/2} \right]\end{aligned}$$

Not surprisingly, this theorem states that if we use the hard thresholding method to regularize the large sample covariance matrices, the consistency rate gets slower when the dependence level ($\mathcal{X}^*(\mathcal{T})$) increases, or in other words, the rate is maximized when $\mathcal{X}^*(\mathcal{T}) = 1$, same as what Bickel and Levina (2008a) shows for the i.i.d case. When $\mathcal{X}^*(\mathcal{T})$ reaches T , it will be offset by T in the denominator. The intuition behind is clear: if dependence is strong, then additional information brought by a “new” observation will be effectively less, i.e. the overall information from T observations will be less correspondingly, which will result in a slower consistency rate. On the other hand, according to the $\log J \mathcal{X}^*(\mathcal{T})/T = o(1)$ requirement, when the dependence level $\mathcal{X}^*(\mathcal{T})$ increases, J must decrease and T must increase to retain the same amount of information.

A very natural question to ask next is: to what extent, the degree of dependence (in terms of β -mixing coefficients) is allowed, while the consistency rate is still the same as the i.i.d. case, i.e. to study the relationship among high dimensionality R , moderate sample size T and β -mixing coefficient β_{mix} .

ASSUMPTION 3.1 $A1 \ \forall t, \mathbb{E} X_{ti} X_{tj} = 0$

$$A2 \ \exists \sigma^2, \forall n, m, m^{-1} \mathbb{E}(X_{ni} X_{nj} + \dots + X_{n+m,i} X_{n+m,j})^2 \leq \sigma^2$$

$$A3 \ \forall t, |X_{ti} X_{tj}| \leq M$$

THEOREM 3.2 (Balance “ J, T, β ” to achieve “good” consistency rate) *Assume the β -mixing sequence $\{X_{ti} X_{tj}\}_{t=1}^T$ satisfies Assumption 3.1 $\forall i, j$ with a high probability. Then, uniformly on $\mathcal{U}_\tau(q, c_0(J), \Sigma)$, for sufficiently large M' also depending on σ^2, M , if $s_T = M'(\sigma^2, M) \sqrt{\frac{\log J}{T}}$, $\log J/T = o(1)$ and the β -mixing coefficient $\beta_{mix} = \mathcal{O}\{(J^{2+\delta'} \sqrt{\log JT})^{-1}\}, \delta' > 0$, we have:*

$$\begin{aligned} \|T_{s_T}(\hat{\Sigma}) - \Sigma\| &= \mathcal{O}_P\left\{c_0(J) \left(\frac{\log J}{T}\right)^{(1-q)/2}\right\} \\ J^{-1} \|T_{s_T}(\hat{\Sigma}) - \Sigma\|_F^2 &= \mathcal{O}_P\left\{c_0(J) \left(\frac{\log J}{T}\right)^{1-q/2}\right\} \end{aligned}$$

As we can see, when dimensionality J increases, since the β -mixing coefficient is controlled by $\mathcal{O}\{(J^{2+\delta'} \sqrt{\log JT})^{-1}\}, \delta' > 0$, the dependence level must decrease at the rate of J^{-2} (skipping the slow varying logs). When J is very large, this means “nearly” independent, which again confirms the result from the previous theorem.

3.2 Choice of Threshold via Cross Validation

Choices of threshold play a fundamental role in implementing this estimation procedure. We choose an optimal threshold by a cross-validation procedure as in Bickel and Levina (2008a) and Bickel and



“Thresholding” Set Ω_1

“Population” Set Ω_2

Figure 3: Illustration of the cross-validation method.

Levina (2008b). In particular, we divide the data set Ω of size T into two consecutive segments, Ω_1 and Ω_2 of size T_1 and T_2 respectively, where T_1 is typically about $T/3$. Then we compare the regularized (via thresholding) “target” quantity $T_s(\hat{\Sigma}_{1,v})$, estimated from Ω_1 , with the “target” quantity $\hat{\Sigma}_{2,v}$, estimated from Ω_2 . Hence $\hat{\Sigma}_{2,v}$ can be viewed as a proxy to the population “target” quantity Σ . The subindex v in $T_s(\hat{\Sigma}_{1,v})$ and $\hat{\Sigma}_{2,v}$ indicates values from the v th split from a total of N repeats. The optimal threshold is then selected as a minimizer (w.r.t. s) of the empirical loss function over N repeats, i.e.

$$\arg \min_s N^{-1} \sum_{v=1}^N \|T_s(\hat{\Sigma}_{1,v}) - \hat{\Sigma}_{2,v}\|_F^2. \quad (4)$$

Similarly, the oracle threshold is then selected as a minimizer w.r.t. s of the oracle loss function over N repeats, i.e.

$$\arg \min_s \mathbb{E} \|T_s(\hat{\Sigma}_{1,v}) - \hat{\Sigma}_{2,v}\|_F^2. \quad (5)$$

Since the data are observed in time, the order of X_t is of importance, and hence a random split of Ω to Ω_1 and Ω_2 is not appropriate in a time series context. Alternatively, we randomly select a consecutive segment of size $T_1 + T_2$ as $\Omega_1 \cup \Omega_2$ from the data set Ω first, and then take the first third of $\Omega_1 \cup \Omega_2$ as Ω_1 ($T_2 \approx 2T_1$) and the remaining two thirds as Ω_2 . Figure 3 provides an illustration for the cross-validation procedure. We repeat this N times as before. Our goal now is to show that the rates of convergence for the empirical loss function and the oracle loss function are of the same order and hence, asymptotically the empirical threshold \hat{s} performs as well as the oracle threshold s_0 selection.

Our theoretical justification is based on adapting the results on the optimal threshold selection in Bickel and Levina (2008a) and the optimal band selection in Bickel and Gel (2011) to a case of optimal choice of a threshold for high dimensional β -mixing time series. Let $W_1, \dots, W_n, \dots, W_{n+B}$ be $J^2 \times 1$ vectors with common mean $\mathbb{E} W$. Let $\|x\|_v = \max_{p=1, \dots, P} |v_p' x|$, $x \in \mathbb{R}^J$, $v_p \in \mathbb{R}^J$, $\|v_p\| = 1$

and $\bar{W}_B = B^{-1} \sum_{p=1}^B W_{n+p}$. Then the empirical and oracle estimates based on W_k are defined as

$$\hat{\mu}^e \stackrel{\text{def}}{=} \arg \min_{p=1, \dots, P} |\bar{W}_B - \hat{\mu}_p|^2 \quad (6)$$

$$\hat{\mu}^o \stackrel{\text{def}}{=} \arg \min_{p=1, \dots, P} |\mathbb{E} W - \hat{\mu}_p|^2 \quad (7)$$

respectively, where $\hat{\mu}_p$ is estimated using W_1, \dots, W_n .

We use Theorem 3 in Bickel and Levina (2008a) as Lemma 3.1 here, which states a result on asymptotic relation between the empirical and oracle estimates $\hat{\mu}^e$ and $\hat{\mu}^o$.

LEMMA 3.1 (Theorem 3 in Bickel and Levina (2008a)) *If the following assumptions (A4, A5, A6) are satisfied*

$$A4 \quad |\hat{\mu}^o - \mathbb{E} W|^2 = \Omega_P(r_n);$$

$$A5 \quad \mathbb{E} \max_{p=1, \dots, P} \|(v_j, W_1 - \mu)\|^2 \leq C\rho(P) \text{ for } v_p \in \mathbb{R}^J, \|v_p\| = 1;$$

$$A6 \quad \rho(P_n) = o(r_n),$$

then we have

$$|\hat{\mu}^e - \bar{W}_B|^2 = |\hat{\mu}^o - \mathbb{E} W|^2 \{1 + o(1)\} = \Omega_P(r_n).$$

Without loss of generality, assume that the number of repeats $N = 1$. Notice that the empirical estimates $T_s(\hat{\Sigma}_{1,v})$ and $\hat{\Sigma}_{2,v}$ play the role of $\hat{\mu}_p$ and \bar{W} here respectively. Hence, if we can verify the conditions of Lemma 3.1, we can apply it to justify the choice of a threshold by cross-validation and show that such regularized covariance matrix of high dimensional time series $\{X_t\}$, with an empirical selected threshold, asymptotically coincides with the regularized estimate selected by oracle. To this end, we also need the auxiliary Lemma 3.2.

LEMMA 3.2 *Assume that v_t is white noise satisfying $\mathbb{E} v_t = 0$, $\mathbb{E} v_t^2 = \sigma^2$ and $\mathbb{E} |v_t|^\beta \leq C < \infty$ for $\beta > 2$. Let $\|V\|_F = 1$. For the β -mixing process satisfying the conditions of Theorem 3.2, we have*

$$\begin{aligned} \mathbb{P} \left(J^{-1} |\text{tr}(V \hat{\Sigma}_B - V \Sigma)| \geq s \right) &\leq K_1 \exp(-K_2 s^2 B) \\ J^{-1} \mathbb{E} \max_{p=1, \dots, P} \left(|\text{tr}\{v_j \hat{\Sigma}_B - \mathbb{E}(v_j \Sigma)\}| \right) &\leq C(q, c_0, M) \sqrt{\log P / B} \end{aligned}$$

with some constants K_1 and K_2 .

THEOREM 3.3 (Consistency of Cross Validation) *Let \hat{s} and s^o be the threshold selected from minimizing the empirical and oracle loss functions (4) and (5) respectively. Then under the conditions of Theorem 3.2 and $\mathcal{O}_P = \Omega_P$, if $B_T = T\varepsilon(T, J)$, $\log P = o\{T^{q/2} c_0(J) J^{-1} (\log J)^{1-q/2} \varepsilon(T, J)\}$, based on Lemma 3.1 and 3.2, then*

$$\|T_{\hat{s}}(\hat{\Sigma}) - \Sigma\|_F = \|T_{s^o}(\hat{\Sigma}) - \Sigma\|_F \{1 + o_P(1)\}.$$

4 The Screen - Cluster - Estimate (SCE) Procedure

To circumvent the problems in semiparametric modeling for high dimensional data with complex spatial structure, in the following three subsections, we state the three-step SCE procedure for constructing and estimating semiparametric models from a large number of unordered explanatory variables with a moderate sample size.

4.1 Screen

- 1 Estimate the $J \times J$ (dependent variable y (x_J) and all explanatory variables x_1, \dots, x_{J-1}) large covariance (Spearman's correlation) matrix using hard thresholding as $T_{\hat{s}}(\hat{\Sigma}) \stackrel{\text{def}}{=} [\tilde{\sigma}_{ij}]$, and only keep and consider the (say K) x 's with nonzero correlation entries with y for following steps.

Without loss of generality, we rename the K x 's as x_1, x_2, \dots, x_K .

Since all observations are standardized first, the previously considered covariance matrix is actually the (Pearson's) correlation coefficient matrix. However, at the "screening" step, we estimate and threshold the large Spearman's rank correlation matrix, where Spearman's rank correlation between x_i and x_j is defined as:

$$\rho_{x_i, x_j} = \frac{\text{Cov}\{F_i(x_i), F_j(x_j)\}}{\sqrt{\text{Var}\{F_i(x_i)\} \text{Var}\{F_j(x_j)\}}}, \quad (8)$$

and F_i and F_j are the cumulative distribution functions of x_i and x_j respectively. It can be seen that the population version of Spearman's rank correlation is just the classic Person's correlation between $F_i(x_i)$ and $F_j(x_j)$. Here we consider Spearman's rank correlation instead of the Pearson's correlation coefficient is because the latter one is sensitive only to a linear relationship between two variables, while the former one is more robust than the Pearson's correlation - that is, more sensitive to nonlinear relationships. It could be viewed as a non-parametric measure of correlation and especially suitable for the non and semiparametric situations we consider here. It assesses how well an arbitrary monotonic function could describe the relationship between two variables. Specifically speaking, it measures the extent to which, as one variable increases, the other variable tends to increase, without requiring that increase to be represented by a linear relationship. If, as the one variable increases, the other decreases, the rank correlation coefficients will be negative. Similar to the consistency results towards the large spatial thresholding covariance (correlation) matrix studied here, Xu and Bickel (2010) established those for the large Spearman's rank correlation matrix (for the i.i.d. case).

At step 1, via hard thresholding, we single out the important predictors by using their Spearman's rank correlations with the response variable y and eliminate all explanatory variables that are "irrelevant" to y . In light of equation (1), we actually get an estimate for $\mathcal{A}_1 \cup \dots \cup \mathcal{A}_S$. Thus we could

reduce the feature space significantly from J to a lower dimensional and more manageable space. Correlation learning is a specific case of independent learning, which ranks the features according to the marginal utility of each feature. The computational expediency and stability are prominently featured in independent learning. This kind of idea is frequently used in applications (Guyon and Elisseeff (2003)) and recently has been carefully studied for its theoretical properties by Fan and Lv (2008) using Pearson’s correlation for variable screening of linear models; Huang et al. (2008) , who proposed the use of marginal bridge estimators to select variables for sparse high dimensional regression models; Fan et al. (2011) using the marginal strength of the marginal nonparametric regression for variable screening of additive models; Hall and Miller (2009) using the generalized correlation for variable selection of linear models.

It is also worthy noticing that the threshold is a global measure (implicitly) depending on all J variables. If we remove some x ’s from the original explanatory variables set, the threshold value will be changed correspondingly. Thus the “relevant” and “irrelevant” regressors will also change.

4.2 Cluster

Motivated by the fact that in a block diagonal matrix, the nonzero entries along the diagonal are denser than those in the off-diagonal region and the assumption w.r.t. equation (1): “ $\forall j \neq l, x_j \in \mathcal{A}_j, x_l \in \mathcal{A}_l, x_j$ and x_l are (conditionally) independent given other x ’s”, we define the following “averaged non-zero” score $S_{\mathcal{A}}$ for a index set \mathcal{A} : $S_{\mathcal{A}} \stackrel{\text{def}}{=} \sum_{i,j \in \mathcal{A}} \mathbf{1}(\tilde{\sigma}_{ij} \neq 0) / |\mathcal{A}|^2$. Here we do not distinguish between the positive and negative values of $\tilde{\sigma}_{ij}$ since they could also be reflected by the corresponding linear coefficients as in equation (1).

- 2 Perform the label permutation procedure for x_1, \dots, x_K to form clusters of (explanatory) variables (or $\mathcal{A}_1, \dots, \mathcal{A}_S$) by utilizing the “averaged non-zero” score $S_{\mathcal{A}}$.
 - 2.1 Rank (in decreasing order) and relabel all $x_1, \dots, x_k, \dots, x_K$ according to $\sum_{1 \leq j \leq K} \mathbf{1}(\tilde{\sigma}_{kj} \neq 0)$ to obtain the “new” x_1, \dots, x_K . Always assume x_1 is in the first block (index set) \mathcal{A}_1 ;
 - 2.2 *Forward* Include x_k ($2 \leq k \leq K$) in the first index set ($x_k \in \mathcal{A}_1$) if $S_{\mathcal{A}_1 \cup \{x_j\}} \geq S_{\mathcal{A}_1}$, and continue searching until the K th variable x_K . Without loss of generality (otherwise just relabel them), we assume $x_1, x_2, \dots, x_{k-1} \in \mathcal{A}_1$.
 - 2.3a (For the case of no overlapping indices among $\mathcal{A}_1, \dots, \mathcal{A}_S$)
Given \mathcal{A}_1 formed in the last step, perform Steps 2.1 and 2.2 again for the variables not in the set \mathcal{A}_1 , i.e. $\{1, 2, \dots, K\} \setminus \mathcal{A}_1$ and construct \mathcal{A}_2 .

2.3b *Backward* (replace Step 2.3a, for the case allowing overlapping indices among $\mathcal{A}_1, \dots, \mathcal{A}_S$)

Given \mathcal{A}_1 , perform Step 2.1 again for the variables not in the set \mathcal{A}_1 , i.e. $\{1, 2, \dots, K\} \setminus \mathcal{A}_1$ and start to construct \mathcal{A}_2 , for example, $x_k \in \mathcal{A}_2$. Let $x_1 \in \mathcal{A}_1 \cap \mathcal{A}_2$ only if $S_{\{x_1\} \cup \mathcal{A}_2} \geq S_{\mathcal{A}_2}$ and continue searching until x_{k-1} . Notice that it is impossible for all $x_1, \dots, x_{k-1} \in \mathcal{A}_2$ because of the way we construct \mathcal{A}_1 in the *forward* step. Continue to construct \mathcal{A}_2 as in the *forward* step by selecting variables w.r.t. $\{1, 2, \dots, K\} \setminus \mathcal{A}_1 \cup \{x_k\}$.

2.4 Continue this procedure until all variables x_1, \dots, x_K have been included into some index set(s) of $\mathcal{A}_1, \dots, \mathcal{A}_S$, where S is the number of selected index sets and $\mathcal{A}_1 \cup \dots \cup \mathcal{A}_S = \{1, 2, \dots, K\}$. Given these, construct the corresponding semiparametric models by equation (1).

At Step 2, if we can permute the variables' labels to have a block diagonal structure for the partition of the consistently estimated covariance matrix, such as the one in Figure 2 (right), we can construct the corresponding class of semiparametric models as specific cases of equation (1). By this step, we are grouping the (explanatory) variables into highly correlated groups, which are, however, weakly correlated with each other. This “independence” property (between the “new” predictors $\beta_s^\top x_{\mathcal{A}_s}, 1 \leq s \leq S$) is actually also required for the groupwise dimension reduction method of Li et al. (2010) for estimation. Except that we use the Spearman's rank correlation instead of the Pearson's correlation for “screening”, a second difference between this work and Fan and Lv (2008) is that we consider the covariance (correlation) matrix for all x_1, \dots, x_{J-1}, y variables instead of just between y and x_1, \dots, x_{J-1} , which is because we need to further group the relevant explanatory variables for semiparametric model construction at the second step.

A very important feature of the proposed label permutation procedure is that it is based on the thresholding regularized covariance matrix instead of the sample one. A related work which tries to discover the ordering of the variables through the metric multi-dimensional scaling method could be found in Wagaman and Levina (2009) for the i.i.d. Gaussian case. Their ultimate goal is to improve covariance matrix estimation rather than order the variables itself. Thus by utilizing the discovered “order” based on the sample covariance matrix, they estimate the large covariance matrix through banding regularization to enjoy the benefits brought by ordering. But in the case of large panels of economic and financial variables as we consider here, our ultimate goal is to cluster the variables to construct the proper semiparametric models instead of “ordering”. For example, in the multiple index model, the order of the first index (or first cluster of variables) and the second index (or second cluster of variables) and the order of variables inside each “cluster” are both unimportant.

This case is also related to the hierarchical clustering, k-means algorithm and correlation clustering problem in computer science (Demaine and Immorlica (2003), Bansal et al. (2004)), which aims to partition a weighted graph with positive and negative edge weights so that negative edges are broken up and positive edges are kept together. However, the correlation clustering algorithm is also based on the sample correlation(s), and has also been shown to be NP-hard. Thus, as a *key* difference with other works in the literature, instead of using the sample covariance (or correlation) matrix for ordering and clustering as Wagaman and Levina (2009), Demaine and Immorlica (2003) and Bansal et al. (2004) did, we implement thresholding regularization for the sample covariance matrix and screening first and then find the corresponding groups through the stepwise label permutation procedure. It is simpler to be implemented than their's, since the thresholding regularized covariance matrix only has limited number of nonzero entries. By doing so, w.r.t. the regression setup, we also simultaneously extract the “relevant” explanatory variables for y (Step 1). Thus, we actually combine *dimension reduction* and *variable clustering*, which is especially suitable for modeling high dimensional data via semiparametric methods.

This procedure is computationally simple for a typical $J \leq 150$ macroeconomic and financial data set since the thresholding regularization procedure removes $J - 1 - K$ “irrelevant” variables first, and then rank the remaining K ones before entering this label permutation procedure. Thus we avoid the NP-hard correlation clustering problem based on the sample covariance matrix.

4.3 Estimate

3 Groupwise dimension reduction with sign constraints.

For Step 3, we implement the groupwise dimension reduction estimation procedure modified from Li et al. (2010). If we implement their method directly, as we can see from Table 1 (details of data presented later), the Spearman’s rank correlations between x_1, \dots, x_K and y are all positive, however, some of their corresponding parametric coefficients are estimated to be *negative* (details presented later in Table 2). This means that the consumer price index negatively depends on them, which is unlikely to be true from an economic point of view. Since we have disjoint groups of variables here and given the meaning of the Spearman’s rank correlation, ideally, the sign of the corresponding parametric coefficient estimate w.r.t. $x_k, 1 \leq k \leq K$ should be the same as the sign of the corresponding Spearman’s rank correlation. This motivates us to add the *sign constraint*, as a refinement, to the groupwise dimension reduction method developed in Li et al. (2010) to secure the *sign consistency*.

Let us first consider a simple linear regression model $\mathbf{E} y = x_1\beta_1 + \dots + x_K\beta_K \stackrel{\text{def}}{=} x^\top \beta$ with the constraint $\beta_1, \dots, \beta_K \geq 0$. The linear coefficient β could be estimated as the minimizer of $\|y - x^\top \beta\|_2^2/2 - \sum_{k=1}^K \lambda_k \beta_k$ with the corresponding nonnegative Lagrange multipliers λ_k 's, $1 \leq k \leq K$. If we denote $\text{diag}[\lambda_1, \dots, \lambda_K]$ by Λ , $\hat{\beta} = (x^\top x)^{-1}(x^\top y + \Lambda) \stackrel{\text{def}}{=} \hat{\beta}_{OLS} + (x^\top x)^{-1}\Lambda$. Intuitively, in case some entry of $\hat{\beta}_{OLS}$, say the k th, is negative, which contradicts the initial requirement $\beta_k \geq 0$, $(x^\top x)^{-1}\lambda_k$ plays the role of adding a positive increment to it, s.t. $\hat{\beta}_k \geq 0$.

Similarly, in our setup, if we use $\tilde{\sigma}_{kJ}$ to denote the Spearman's rank correlation estimate between x_k and y (x_J) extracted from $T_{\hat{s}}(\hat{\Sigma})$ and add the sign constraint $\text{sign}(\tilde{\sigma}_{kJ})\beta_k \geq 0$ to the estimation procedure of Li et al. (2010) (β_s here corresponds to their β_g), a simple calculation shows that we just need replace their estimation equation (15) for $\beta \stackrel{\text{def}}{=} (\beta_1, \dots, \beta_K)^\top$ by (β here corresponds to their ζ):

$$\begin{aligned} \hat{\beta} &= \left\{ \sum_{i=1}^T \sum_{j=1}^T \mathbf{R}^{ij} \mathbf{R}^{ij\top} K_h(V^j - V^i) \right\}^{-1} \left\{ \sum_{i=1}^n \sum_{j=1}^n (Y^j - a^i) K_h(V^j - V^i) \mathbf{R}^{ij} + \Lambda' \right\}, \\ &= \hat{\zeta} + \left\{ \sum_{i=1}^T \sum_{j=1}^T \mathbf{R}^{ij} \mathbf{R}^{ij\top} K_h(V^j - V^i) \right\}^{-1} \Lambda' \end{aligned} \quad (9)$$

where Λ' is a $K \times K$ diagonal matrix $\text{diag}[\lambda_1 \text{sign}(\tilde{\sigma}_{1J}), \dots, \lambda_K \text{sign}(\tilde{\sigma}_{KJ})]$; $\{\lambda_k, 1 \leq k \leq K\}$ are the hyperparameters; $\hat{\zeta}$ and other variables are the same as in equation (15) of Li et al. (2010). In general, selection of λ 's requires minimizing some loss function. Motivated by the discussion above for the simple linear regression case, when $\text{sign}(\hat{\zeta}_k)$ is the same as $\text{sign}(\tilde{\sigma}_{kJ})$, we simply choose $\lambda_k = 0$, otherwise choose λ_k to be the minimum (positive) value s.t. $\hat{\beta}_k = 0$. By our experience, this works well and the convergence of the iterative estimation procedure is achieved within 19 iteration steps (10^{-6} as the tolerance) for modeling CPI, which is to be presented in Section 5. Then by the property of the convex minimization problem, if a local minimum exists, it is also a global minimum.

Overall, similar to Fan and Lv (2008)'s "screen first; fit later" approach for modeling high dimensional data, ours could be considered as the "screen first; group second; fit third" approach. Alternatively, Bickel et al. (2009) and Meinshausen and Bühlmann (2006) consider the "fit first; screen later" approach. In general, a great deal of work is needed to compare "screen first; fit later" type of methods with "fit first; screen later" types of method in terms of consistency and oracle properties. But when the spatial structure is complex (thus we need deviate from linearity), in terms of semiparametric modeling, as we have discussed in Section 1, the later one might face several main limitations, while ours, together with the estimation method modified from Li et al. (2010), as a special case of the former one, could circumvent these issues and would be faster when dealing with higher dimensionality.

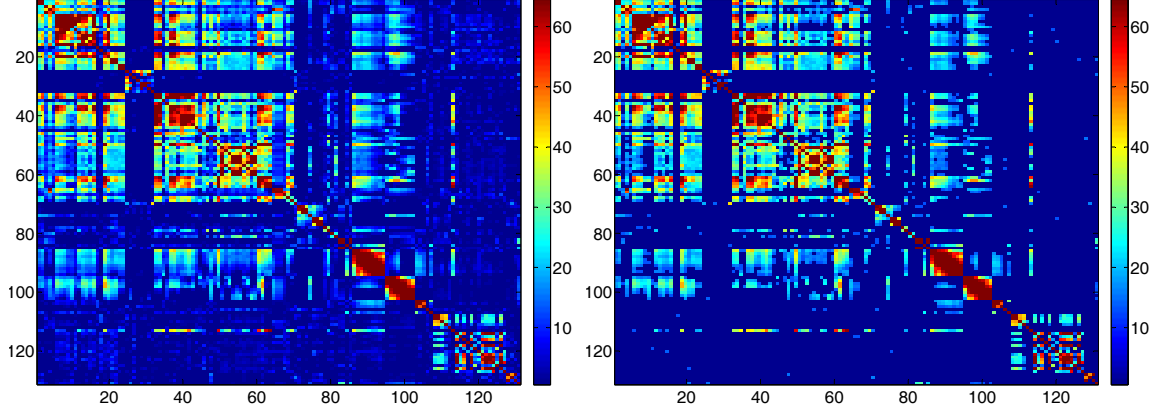


Figure 4: Sample and regularized covariance matrices (after multiplying each entry’s value by 100).

5 Application

We use the dataset of Stock and Watson (2005). This dataset contains 131 monthly macro indicators covering a broad range of categories including income, industrial production, capacity, employment and unemployment, consumer prices, producer prices, wages, housing starts, inventories and orders, stock prices, interest rates for different maturities, exchange rates, money aggregates and so on. The time span is from January 1959 to December 2003. We apply logarithms to most of the series except those already expressed in rates. The series are transformed to obtain stationarity by taking (the 1st or 2nd order) differences of the raw data series (or the logarithm of the raw series). Then all observations are standardized.

CPI	0.62	0.76	0.65	0.65	0.51	0.67	0.65	0.25	0.28	0.28	0.27	0.17	0.25	0.14	0.25
121	0.61	0.55	0.51	0.56	0.52	0.51	0.17	0.22	0.27	0.19	0.20	0.22	0.14	0.32	
123		0.66	0.62	0.47	0.73	0.65	0.24	0.26	0.26	0.25	0.19	0.22	0.13	0.25	
122			0.65	0.55	0.70	0.71	0.29	0.28	0.24	0.23	0.13	0.21	0	0	
124				0.45	0.63	0.80	0.23	0.25	0.26	0.30	0	0.58	0.26	0	
116					0.54	0.51	0.20	0.25	0.24	0	0.19	0	0.15	0	
118						0.77	0.29	0.31	0.29	0.27	0.19	0	0	0	
126							0.31	0.35	0.30	0.34	0	0.14	0	0	
108								0.87	0.46	0	0	0	0	0	
109									0.47	0	0	0	0	0	
110										0	0	0	0	0	
115											0	0	0	0	
119												0	0.35	0	
127													0	0.27	
125														0	
120															

Table 1: Partition of the $T_s(\hat{\Sigma}_{1,v})$ w.r.t. *CPI* and the 15 “relevant” variables and the diagonal entries denote the indices (in the original data set) of the corresponding variables.

Variable	Meaning	Coefficients	Coefficients
$PWFCSA_{108}$	Producer Price Index: Finished Goods	0.150	0.207
$PWIMSA_{109}$	PPI: Finished Consumer Goods	0.741	0.763
$PWCMSA_{110}$	PPI: Intermed. Mat. Supplies & Components	0.074	0.196
$PU84_{116}$	CPI-U: Transportation	0.307	0.304
PUC_{118}	CPI-U: Commodities	0.290	0.235
$PUXF_{121}$	CPI-U: All Items Less Food	0.397	0.378
$PUXHS_{122}$	CPI-U: All Items Less Shelter	-0.034	0
$PUXM_{123}$	CPI-U: All Items Less Medical Care	0.246	0.217
$GMDC_{124}$	PCE,IMPL PR DEFL:PCE	-0.146	0
$GMDCN_{126}$	PCE,IMPL PR DEFL:PCE; Nondurables	-0.061	0
$PUCD_{119}$	CPI-U: Durables	-0.639	0.635
$GMDCD_{125}$	PCE,IMPL PR DEFL:PCE; Durables	-0.770	0.772
PUS_{120}	CPI-U: Services	-0.994	0.979
$GMDCS_{127}$	PCE,IMPL PR DEFL:PCE; Services	0.111	0.203
$PU83_{115}$	CPI-U: Apparel & Upkeep	1	1

Table 2: Detailed meanings of the variables with the corresponding parametric coefficients' estimates using the groupwise dimension reduction method without (3rd column) and with (4th column) sign constraints.

Figure 4 contains plots of the sample and thresholding regularized Spearman's rank correlation matrices based on the "optimal" threshold 0.13 selected by the cross validation procedure discussed in subsection 3.2 with $T_1 = 120, T_2 = 240$. The variables of special interest include the consumer price index (CPI) as a measure of prices and an economic indicator. The annual percentage change in CPI is used as a measure of inflation. CPI can be used to index (i.e., adjust for the effect of inflation) the real value of wages, salaries and pensions, and also for regulating prices and deflating monetary magnitudes to show changes in real values. Besides being a deflator of other economic series, it is also a means of adjusting dollar values. Thus CPI is one of the most closely watched national economic statistics. To this end, we use modeling of CPI to illustrate our method. Table 1 displays the partition of the $T_s(\hat{\Sigma}_{1,v})$ w.r.t. the variables relevant to CPI. The detailed meanings of these variables (and corresponding three digits indices in the data set provided by Stock and Watson (2005)) are given in Table 2. By the forward (and backward) procedure discussed in Section 4, we find the following index sets for constructing semiparametric models for modeling CPI. The one without overlapping is

presented on the LHS, and that allowing overlapping is presented on the RHS.

$$\begin{array}{ll}
\mathcal{A}_1 = 108 - 110, 116, 118, 121 - 124, 126 & \mathcal{A}'_1 = 108 - 110, 116, 118, 121 - 124, 126 \\
\mathcal{A}_2 = 119, 125 & \mathcal{A}'_2 = 121 - 124, 115 \\
\mathcal{A}_3 = 120, 127 & \mathcal{A}'_3 = 121 - 123, 119 \\
\mathcal{A}_4 = 115 & \mathcal{A}'_4 = 121 - 123, 127 \\
& \mathcal{A}'_5 = 121, 123, 125 \\
& \mathcal{A}'_6 = 120, 121, 123
\end{array}$$

Notice that $\mathcal{A}'_1 = \mathcal{A}_1$, and the main difference between these two methods comes from $\mathcal{A}_2 - \mathcal{A}_4$ and $\mathcal{A}'_2 - \mathcal{A}'_6$, i.e. how to allocate the 119, 125, 120, 127, 115th variables, which originally results from the rank correlations between them and the 121 – 124th variables. As we can see from Table 2, the 121 – 124th variables are very close to the y variable: CPI-U: All Items (82-84=100, SA) except one item (food, shelter or medical care) or the implicit price deflator (of personal consumption expenditures).

Due to the identification and estimation problems we discussed before, from now on, we mainly concentrate on the disjoint index sets case and suggest the following semiparametric model for modeling CPI: $E(CPI_{114}) =$

$$\begin{aligned}
& g_1 \left(\beta_{108} PWFC SA_{108} + \beta_{109} PWIM SA_{109} + \beta_{110} PWCMSA_{110} + \beta_{116} PU84_{116} + \beta_{118} PUC_{118} \right. \\
& \quad \left. + \beta_{121} PUXF_{121} + \beta_{122} PUXHS_{122} + \beta_{123} PUXM_{123} + \beta_{124} GMDC_{124} + \beta_{126} GMDCN_{126} \right) \\
& \quad + g_2 \left(\beta_{119} PUCD_{119} + \beta_{125} GMDCD_{125} \right) + g_3 \left(\beta_{120} PUS_{120} + \beta_{127} GMDCS_{127} \right) + g_4 \left(PU83_{115} \right), \quad (10)
\end{aligned}$$

where g_1, \dots, g_4 are the unknown link functions to be estimated nonparametrically and $\beta_{108}, \dots, \beta_{127}$ are unknown parameters which belong to the parameter space. The variables $PUCD_{119}$ and $GMDCD_{125}$ denote the consumer price index and implicit price deflator (of personal consumption expenditures) for durable goods respectively. Thus \mathcal{A}_2 could be interpreted as the index set for durable goods. Very similarly, \mathcal{A}_3 and \mathcal{A}_4 could be interpreted as the index sets for service, and apparel and upkeep respectively which are also important factors affecting consumer price index. All common factors strongly associated with CPI are included in \mathcal{A}_1 . Compared with the linear, additive or single index models, the model (10) actually combines flexibility in statistical modeling and interpretability from an economic point of view, while being kept close to the data's complex spatial structure.

We further employ the groupwise dimension reduction method in Li et al. (2010) to estimate (10) and present the parametric coefficients' estimate in Table 1 (3rd column). Contradicting to

the background knowledge of economics and the positive Spearman's rank correlations between CPI and $PUXHS_{122}$, $GMDC_{124}$, $GMDCN_{126}$, $PUCD_{119}$, $GMDCD_{125}$, PUS_{120} shown in Table 2, their corresponding parametric coefficients are estimated to be *negative*. This means that the consumer price index negatively depends on them, which is unlikely to be true. Finally we apply the modified procedure with sign constraints to estimate (10) again and present the corresponding parametric coefficients' estimates in the last column of Table 2 with the explained variation 85.8%. While β_{119} , β_{125} and β_{120} are estimated positively, β_{122} , β_{124} and β_{126} are estimated to be 0, which means $PUXHS_{122}$, $GMDC_{124}$ and $GMDCN_{126}$ could be eliminated from the model. To compare with the linear models and see the advantages of semiparametrics, we also consider the linear model using all other 130 variables (except CPI itself). The explained variation w.r.t. the LARS estimate (least angle regression, developed by Efron et al. (2004)) is 80.4%.

Besides the measure of prices, other variables of special interest include a measure of real economic activity and a monetary policy instrument. As in Christiano et al. (1999), we use employment as an indicator of real economic activity measured by the number of employees on non-farm payrolls (EMPL). The monetary policy instrument is the Federal Funds Rate (FFR). If we apply the SCE approach to estimate EMPL and FFR, the explained variation is 99.9% and 97.6% respectively, while the corresponding LARS estimates' is 99.6% and 86.7%. These results are summarized in Table 3. Thus we see that we could reduce the SSE approximately by 27.4% for CPI and 82.0% for FFR through considering the (flexible and proper) semiparametrics. The improvement for EMPL is not significant since the LARS estimate has already performed quite well.

	CPI	EMPL	FFR
R^2 (SCE)	85.8%	99.9%	97.6%
R^2 (LARS)	80.4%	99.6%	86.7%

Table 3: Explained variation of the SCE and LARS estimates for CPI, EMPL and FFR.

6 Concluding Remarks and Discussions

In this paper, we consider estimating a large spatial covariance matrix of the generalized m dependent and β -mixing time series (with J variables and T observations) by hard thresholding regularization. We quantify the interplay between the estimators' consistency rate and the time dependence level, discuss an intuitive resampling scheme for threshold selection, and prove a general cross-validation result that justifies this approach. Given a consistently estimated large sparse covariance matrix,

by utilizing the natural links among graphical models, semiparametrics and large spatial covariance matrix, we propose a novel forward (and backward) label permutation procedure to form a block diagonal structure for it and construct the corresponding low dimensional semiparametric model. Finally we apply this method to study the spatial structure of large panels of economic and financial time series to find the proper semiparametric structure for estimating the consumer price index (CPI) and present its superiority over the linear models.

Choice of Threshold

Concerning the choice of threshold in the context of time series analysis, if we are mainly targeting estimation performance of the corresponding semiparametric models instead of minimizing the loss functions (6) and (7) related to the covariance matrix estimation, we might directly consider minimizing the estimation error based on the selected semiparametric model, for example (2), s.t. the prediction performance might be optimized.

Other Measures of Dependence for Screening

The information given by a Pearson's correlation coefficient is not enough to define the dependence structure between random variables. Except the Spearman's rank correlation we used here, distance correlation, Székely et al. (2007) and Brownian covariance (correlation), Székely and Rizzo (2009) were also introduced to address the deficiency of Pearson's correlation that it can be zero for dependent random variables; zero distance correlation and zero Brownian correlation imply independence. The correlation ratio is able to detect almost any functional dependency, and the entropy-based mutual information/total correlation is capable of detecting even more general dependencies. We want to point out that the Step 1 of the SCE procedure could be very easily extended to these measures above and the threshold value could be selected by the cross-validation procedure similarly.

It is also noteworthy that Fan et al. (2011) considers the independence screening procedure by ranking the explanatory variable's importance according to the descent order of the residual sum of squares of the componentwise nonparametric regressions or the marginal strength of the marginal nonparametric regression. By doing that, they (implicitly) assume that the true semiparametric structure is additive, which is different from our ultimate goal here: construct the proper semiparametric structure.

Theoretical Study of the Screening Step

Noticing that $F_i(x_i)$ and $F_j(x_j)$ in the formula (8) follow the uniform distribution on $[0, 1]$, thus (8) could be simplified as $\rho_{x_i, x_j} = 12 \mathbf{E}\{F_i(x_i)F_j(x_j)\} - 3$. Similar to the "sure independence screening" property of Fan and Lv (2008) using Pearson's correlation for variable screening of linear models, to study the theoretical property of the screening step here based on the Spearman's rank correlation,

parallel to the equation (20) “ $\omega = X^\top y = X^\top X\beta + X^\top \varepsilon$ ” of Fan and Lv (2008), we could define

$$\begin{aligned}\omega &= (\omega_1, \dots, \omega_{J-1}) \\ \omega_j &= 12F_j(x_j)F_J(x_J) - 3 \stackrel{\text{def}}{=} 12F_j(x_j)F_y(y) - 3, \quad 1 \leq j \leq J-1 \\ &= 12F_j(x_j)F_y \left\{ \sum_{s=1}^S g_s(\beta_s^\top x_{\mathcal{A}_s}) + \varepsilon_0 \right\} - 3,\end{aligned}\tag{11}$$

where ε_0 is the (conditional) mean-zero error term from approximating y by $\sum_{s=1}^S g_s(\beta_s^\top x_{\mathcal{A}_s})$ in (1). Similar to the idea of the (group) MAVE method of Xia et al. (2002), Li et al. (2010), we notice that

$$\partial g_s(\beta_s^\top x_{\mathcal{A}_s}) / \partial x_{\mathcal{A}_s} = g'_s(\beta_s^\top x_{\mathcal{A}_s}) \beta_s,$$

provided by $g'_s(\beta_s^\top x_{\mathcal{A}_s})$ is well defined. Thus applying the Taylor expansion to $\sum_{s=1}^S g_s(\beta_s^\top x_{\mathcal{A}_s}) + \varepsilon_0$ at x' will help linearize it as:

$$\begin{aligned}a &+ \sum_{s=1}^S g'_s(\beta_s^\top x'_{\mathcal{A}_s}) \beta_s^\top (x - x')_{\mathcal{A}_s} + \mathcal{O} \left\{ \sum_{s=1}^S (x - x')_{\mathcal{A}_s}^\top (x - x')_{\mathcal{A}_s} \right\} + \varepsilon_0 \\ &\stackrel{\text{def}}{=} a + \sum_{s=1}^S b_s \beta_s^\top (x - x')_{\mathcal{A}_s} + \varepsilon.\end{aligned}$$

Therefore, we could rewrite (11) as

$$12F_j(x_j)F_y \left\{ a + \sum_{s=1}^S b_s \beta_s^\top (x - x')_{\mathcal{A}_s} + \varepsilon \right\} - 3.\tag{12}$$

Studying the property of (12) will be the main focus. However, due to the presence of the cumulative density functions F_j and F_y here, this is expected to be much more complex than the Pearson's correlation case. We hope that the other people could further investigate this.

Acknowledgement This work was partially motivated during a conversation with Prof Lixing Zhu in Hong Kong in Feb, 2010. The author is very grateful to Prof Peter Bickel, Prof Lixing Zhu, Prof Lexin Li, Mu Cai and Ying Xu for very interesting discussions and comments on this and related topics. In particular, I would like to thank Prof Peter Bickel for sponsoring my stay at the University of California, Berkeley.

7 Appendix

Proof of Theorem 3.1 The proof of this theorem is based on the ones of Theorem 1 and 2 in Bickel and Levina (2008a) up to a modification of the bound on $P\{\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| \geq s\}$, as remarked by their

subsection 2.3. By the definition of $\hat{\sigma}_{ij}$ in (3) and the assumption that for all i and j , $|X_{ti}X_{tj}| \leq M_t$ holds with a high probability, applying the (extended) Mcdiarmid inequality, see Theorem 2.1 of Janson (2004), to the sum of dependent random vectors $\sum_{t=1}^T |X_{ti}X_{tj}|$ yields:

$$\mathbb{P}\{\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| \geq s\} \leq J^2 \exp \left\{ -\frac{s^2 T^2}{\mathcal{X}^*(\mathcal{T}) \sum_t M_t^2} \right\} = \exp\{(2 - M'^2) \log J\},$$

where $s_T = M' \sqrt{\log J \mathcal{X}^*(\mathcal{T})/T}$ with sufficiently large M' also depending on C' with $\sum_{t=1}^T M_t^2 C'/T \leq C'$ and $\log J \mathcal{X}^*(\mathcal{T})/T = o(1)$. Since equation (10) in Bickel and Levina (2008a) holds, others go through verbatim. This completes the proof. \square

Proof of Theorem 3.2 The proof of this theorem is also based on the ones of Theorem 1 and 2 in Bickel and Levina (2008a) up to a modification of the bound on $\mathbb{P}\{\max_{i,j} |\hat{\sigma}_{ij} - \sigma_{ij}| \geq s\}$. Assume the β -mixing sequence $\{X_{ti}X_{tj}\}_{t=1}^T$ to satisfy Assumption 3.1 $\forall i, j$, applying the Bernstein type inequality for β -mixing random variables $\{X_{ti}X_{tj}\}_{t=1}^T$, see Theorem 4 of Doukhan (1994)[P.36], yields that, $\forall \varepsilon > 0$ ($\theta \stackrel{\text{def}}{=} \varepsilon^2/4$) and $\forall 0 < q \leq 1$,

$$\mathbb{P}(|\sum_{t=1}^T X_{ti}X_{tj}| \geq sT) \leq \underbrace{4 \exp \left[-\frac{(1-\varepsilon)3(1+\theta)s^2 T}{2\{3(1+\theta)\sigma^2 + qMsT\}} \right]}_{\stackrel{\text{def}}{=} A} + \underbrace{2 \frac{(1+\theta)\beta_{mix}}{q}}_{\stackrel{\text{def}}{=} B}. \quad (13)$$

To make $J^2(A+B)$ arbitrarily small, we choose $s_T = M' \sqrt{\frac{\log J}{T}}$ with sufficiently large M' also depending on $\varepsilon, \sigma^2, M, \log J/T = o(1)$, $q = 3(1+\theta)\sigma^2/(MsT)$, and $\beta_{mix} = \mathcal{O}\{(J^{2+\delta'} \sqrt{\log JT})^{-1}\}$ with $\delta' > 0$. Thus A and B are bounded by $\exp(-M'^2 \log J)$ and $J^{-(2+\delta')}$ respectively, which can be arbitrarily close to 0. This completes the proof. \square

Proof of Lemma 3.2 Since

$$\mathbb{P}\left(J^{-1}|tr(V\hat{\Sigma}_B - V\Sigma)| \geq s\right) \leq \mathbb{P}\left(J^{-1}|tr(B^{-1}\sum_{p=1}^B V X_p X_p^\top - V\Sigma)| \geq s\right),$$

and $tr(X_p X_p^\top) = tr(X_p^\top X_p)$, $tr(\sum_{p=1}^B V X_p X_p^\top - V\Sigma) \leq J \sum_{p=1}^B \bar{X}_p^2$ with $\bar{X}_p = J^{-1} \sum_{j=1}^J X_{pj}$, applying the same inequality as in (13) to $\sum_{p=1}^B \bar{X}_p^2$ leads to

$$\mathbb{P}\left(J^{-1}|tr(V\hat{\Sigma}_B - V\Sigma)| \geq s\right) \leq K_1 \exp(-K_2 s^2 B)$$

with some constants K_1 and K_2 .

Consequently we also have

$$\mathbb{P}\left(J^{-1} \max_{p=1, \dots, P} |tr(V_p \hat{\Sigma}_B - V_p \Sigma)| \geq s\right) \leq \mathbf{1}(0 \leq s \leq x) + K_1 P \exp(-K_2 s^2 B) \mathbf{1}(s > x). \quad (14)$$

If we integrate (14), i.e.

$$J^{-1} \mathbb{E} \max_{p=1, \dots, P} \left(|tr\{v_j \hat{\Sigma}_B - \mathbb{E}(v_j \Sigma)\}| \right) \leq x + K_1 P \int_x^\infty \exp(-K_2 s^2 B) ds, \quad (15)$$

and minimize the RHS of (15) over x as $P \rightarrow \infty$, we find that the minimizer satisfies $x = C(q, c_0, M) \sqrt{\log P/B} \{1 + o(1)\}$. Hence

$$J^{-1} \mathbb{E} \max_{p=1, \dots, P} \left(|tr\{v_j \hat{\Sigma}_B - \mathbb{E}(v_j \Sigma)\}| \right) \leq C(q, c_0, M) \sqrt{\log P/B}. \quad \square$$

Proof of Theorem 3.3 Based on Lemma 3.2, we conclude that $\rho(P)$ from the second condition of Lemma 3.1 satisfies

$$\rho(P) \leq C(q, c_0, M) J^2 \log P/B.$$

Hence, Lemma 3.2 implies that

$$\mathbb{E} \|B^{-1} \sum_{p=1}^B X_p X_p^\top - \Sigma\|_v \leq C_1 \rho(P).$$

Hence, if we select $B_T = T\varepsilon(T, J)$ and $\log P = o\{T^{q/2} c_0(J) J^{-1} (\log J)^{1-q/2} \varepsilon(T, J)\}$, the conditions of Lemma 3.1 are satisfied and Theorem 3.3 follows. \square

References

- Ahn, H. and Powell, J. L. (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. Journal of Econometrics, 58(1-2):3–29.
- Bansal, N., Blum, A., and Chawla, S. (2004). Correlation clustering. Machine Learning, 56:89–113.
- Bickel, P. and Gel, Y. (2011). Banded regularization of autocovariance matrices in application to parameter estimation and forecasting of time series. Journal of the Royal Statistical Society, Series B, forthcoming.
- Bickel, P. and Levina, E. (2008a). Covariance regularization by thresholding. Ann. Statist., 36(6):2577–2604.
- Bickel, P. and Levina, E. (2008b). Regularized estimation of large covariance matrices. Ann. Statist., 36(1):199–227.
- Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. Annals of Statistics, 37(4):1705–1732.

- Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Heidelberg: Springer Verlag.
- Cai, T. and Zhou, H. (2011). Optimal rates of convergence for sparse covariance matrix estimation. Annals of Statistics, forthcoming.
- Carroll, R. J., Fan, J., Gijbels, I., and Wand, M. P. (1995). Generalized partially linear single-index models. Journal of the American Statistical Association, 92:477–489.
- Christiano, L., Eichenbaum, M., and Evans, C. (1999). Monetary policy shocks: What have we learned and to what end? Handbook of Macroeconomics, 1(1):65–148.
- Demaine, E. D. and Immorlica, N. (2003). Correlation clustering with partial information. In In Proceedings of the 6th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems, pages 1–13. Springer.
- Doukhan, P. (1994). Mixing: Properties and Examples. Heidelberg: Springer Verlag, 1994.
- Efron, B., Hastie, T., Johnstone, L., and Tibshirani, R. (2004). Least angle regression. Annals of Statistics, 32:407–499.
- El Karoui, N. (2008). Operator norm consistent estimation of large-dimensional sparse covariance matrices. Ann. Statist., 36(6):2717–2756.
- Fan, J., Fan, Y., and Lv, J. (2008). High dimensional covariance matrix estimation using a factor model. Journal of Econometrics, 147(1):186–197.
- Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. Journal of the American Statistical Association, 0(0):1–14.
- Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal Of The Royal Statistical Society Series B, 70(5):849–911.
- Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3:1157–1182.
- Hall, P. and Miller, H. (2009). Using generalized correlation to effect variable selection in very high dimensional problems. Journal of Computational and Graphical Statistics, 18(3):533–550.
- Hastie, T. and Tibshirani, R. (1990). Generalized Additive Models. Chapman & Hall/CRC.

- Horowitz, J. (1998). Semiparametric methods in econometrics. Number 131 in Lecture notes in statistics. Springer, New York, NY.
- Huang, J., Horowitz, J. L., and Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist., 36(2):587–613.
- Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. Annals of Statistics, 38:2282–2313.
- Huang, J. and Zhang, T. (2009). The Benefit of Group Sparsity. ArXiv e-prints.
- Ichimura, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. Journal of Econometrics, 21(157-178).
- Ichimura, H. and Lee, L. F. (1991). Semiparametric least squares estimation of multiple index models: Single equation estimation. In Nonparametric and Semiparametric Methods in Econometrics and Statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics, eds. W. A. Barnett, J. Powell, and G. Tauchen, Cambridge, U.K.: Cambridge University Press.
- Janson, S. (2004). Large deviations for sums of partly dependent random variables. Random Structures Algorithms, 24(3):234–248.
- Li, L., Li, B., and Zhu, L. (2010). Groupwise dimension reduction. Journal of the American Statistical Association, 105(491):1188–1201.
- Marčenko, V. A. and Pastur, L. A. (1967). Distributions of eigenvalues of some sets of random matrices. Mathematics of the USSR-Sbornik, 1(4):457.
- Meier, L., van de Geer, S., and Bühlmann, P. (2009). High-dimensional additive modeling. Ann. Statist., 37(6B):3779–3821.
- Meinshausen, N. and Bühlmann, P. (2006). High dimensional graphs and variable selection with the lasso. Annals of Statistics, 34:1436–1462.
- Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009). Sparse additive models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 71(5):1009–1030.
- Song, S. and Bickel, P. (2011). Large vector auto regressions. submitted.

- Speckman, P. (1988). Kernel smoothing in partial linear models. Journal of the Royal Statistical Society. Series B (Methodological), 50(3):pp. 413–436.
- Stock, J. H. and Watson, M. W. (2005). An empirical comparison of methods for forecasting using many predictors. Manuscript, Princeton University.
- Stoker, T. M. (1986). Consistent estimation of scaled coefficients. Econometrica, 54(6):1461–1481.
- Székely, G. J. and Rizzo, M. L. (2009). Brownian distance covariance. Ann. Appl. Stat., 3(4):1236–1265.
- Szkely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. The Annals of Statistics, 35(6):pp. 2769–2794.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B, 58(1):267–288.
- Volkonskii, V. A. and Rozanov, Y. A. (1959). Some limit theorems for random functions. part i. Theory of Probability and its Applications, 4(2):178–197.
- Wagaman, A. and Levina, E. (2009). Discovering sparse covariance structures with the isomap. Journal of Computational and Graphical Statistics, 18:551–572.
- Wu, W. B. and Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. Biometrika, 90(4):pp. 831–844.
- Xia, Y. (2008). A multiple-index model and dimension reduction. Journal of the American Statistical Association, 103(484):1631–1640.
- Xia, Y., Tong, H., Li, W., and Zhu, L. (2002). An adaptive estimation of dimension reduction space. J. Roy. Statist. Soc. B, 64:363–410.
- Xiao, H. and Wu, W. (2011). Covariance matrix estimation for stationary time series. Preprint.
- Xu, Y. and Bickel, P. (2010). Rank correlation matrices. Working paper.
- Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. Journal of the American Statistical Association, 97:1042–1054.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B, 68(1):49–67.